

Proximal Newton-type methods for minimizing composite functions

Jason D. Lee^{*†} Yuekai Sun^{*†} Michael A. Saunders[‡]

May 31, 2013

Abstract

We generalize Newton-type methods for minimizing smooth functions to handle a sum of two convex functions: a smooth function and a non-smooth function with a simple proximal mapping. We show that the resulting proximal Newton-type methods inherit the desirable convergence behavior of Newton-type methods for minimizing smooth functions, even when search directions are computed inexactly. Many popular methods tailored to problems arising in bioinformatics, signal processing, and statistical learning are special cases of proximal Newton-type methods, and our analysis yields new convergence results for some of these methods.

Technical report no. SOL 2013-1

Department of Management Science and Engineering, Stanford University, Stanford, California, May 31, 2013.

1 Introduction

Many problems of relevance in bioinformatics, signal processing, and statistical learning can be formulated as minimizing a *composite function*:

$$\underset{x \in \mathbf{R}^n}{\text{minimize}} \quad f(x) := g(x) + h(x), \quad (1.1)$$

where g is a convex, continuously differentiable loss function, and h is a convex but not necessarily differentiable penalty function or regularizer. Such problems include the *lasso* [23], the *graphical lasso* [10], and trace-norm matrix completion [5].

We describe a family of Newton-type methods for minimizing composite functions that achieve superlinear rates of convergence subject to standard assumptions. The methods can be interpreted as generalizations of the classic

^{*}J. Lee and Y. Sun contributed equally to this work.

[†]Institute for Computational and Mathematical Engineering, Stanford University, Stanford, California.

[‡]Department of Management Science and Engineering, Stanford University, Stanford, California.

proximal gradient method that account for the curvature of the function when selecting a search direction. Many popular methods for minimizing composite functions are special cases of these *proximal Newton-type methods*, and our analysis yields new convergence results for some of these methods.

1.1 Notation

The methods we consider are *line search methods*, which means that they produce a sequence of points $\{x_k\}$ according to

$$x_{k+1} = x_k + t_k \Delta x_k,$$

where t_k is a *step length* and Δx_k is a *search direction*. When we focus on one iteration of an algorithm, we drop the subscripts (e.g. $x_+ = x + t\Delta x$). All the methods we consider compute search directions by minimizing local quadratic models of the composite function f . We use an accent $\hat{\cdot}$ to denote these local quadratic models (e.g. \hat{f}_k is a local quadratic model of f at the k -th step).

1.2 First-order methods

The most popular methods for minimizing composite functions are *first-order methods* that use *proximal mappings* to handle the nonsmooth part h . SpaRSA [26] is a popular *spectral projected gradient* method that uses a *spectral step length* together with a *nonmonotone line search* to improve convergence. TRIP [13] also uses a spectral step length but selects search directions using a trust-region strategy.

We can accelerate the convergence of first-order methods using ideas due to Nesterov [15]. This yields *accelerated first-order methods*, which achieve ϵ -suboptimality within $O(1/\sqrt{\epsilon})$ iterations [24]. The most popular method in this family is the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [1]. These methods have been implemented in the package TFOCS [3] and used to solve problems that commonly arise in statistics, signal processing, and statistical learning.

1.3 Newton-type methods

There are two classes of methods that generalize Newton-type methods for minimizing smooth functions to handle composite functions (1.1). *Nonsmooth Newton-type methods* [27] successively minimize a local quadratic model of the composite function f :

$$\hat{f}_k(y) = f(x_k) + \sup_{z \in \partial f(x_k)} z^T (y - x_k) + \frac{1}{2} (y - x_k)^T H_k (y - x_k),$$

where H_k accounts for the curvature of f . (Although computing this Δx_k is generally not practical, we can exploit the special structure of f in many statistical learning problems.) Our proximal Newton-type methods approximates

only the smooth part g with a local quadratic model:

$$\hat{f}_k(y) = g(x_k) + \nabla g(x_k)^T(y - x_k) + \frac{1}{2}(y - x_k)^T H_k(y - x_k) + h(y).$$

where H_k is an approximation to $\nabla^2 g(x_k)$. This idea can be traced back to the *generalized proximal point method* of Fukushima and Miné [11]. Many popular methods for minimizing composite functions are special cases of proximal Newton-type methods. Methods tailored to a specific problem include **glmnet** [9], **LIBLINEAR** [28], **QUIC** [12], and the Newton-LASSO method [16]. Generic methods include *projected Newton-type methods* [22, 21], proximal quasi-Newton methods [20, 2], and the method of Tseng and Yun [25, 14].

There is a rich literature on solving generalized equations, monotone inclusions, and variational inequalities. Minimizing composite functions is a special case of solving these problems, and proximal Newton-type methods are special cases of Newton-type methods for these problems [17]. We refer to [18] for a unified treatment of *descent methods* (including proximal Newton-type methods) for such problems.

2 Proximal Newton-type methods

We seek to minimize composite functions $f(x) := g(x) + h(x)$ as in (1.1). We assume g and h are closed, convex functions. g is continuously differentiable, and its gradient ∇g is Lipschitz continuous. h is not necessarily everywhere differentiable, but its *proximal mapping* (2.1) can be evaluated efficiently. We refer to g as “the smooth part” and h as “the nonsmooth part”. We assume the optimal value f^* is attained at some optimal solution x^* , not necessarily unique.

2.1 The proximal gradient method

The proximal mapping of a convex function h at x is

$$\text{prox}_h(x) := \arg \min_{y \in \mathbf{R}^n} h(y) + \frac{1}{2} \|y - x\|^2. \quad (2.1)$$

Proximal mappings can be interpreted as generalized projections because if h is the indicator function of a convex set, then $\text{prox}_h(x)$ is the projection of x onto the set. If h is the ℓ_1 norm and t is a step-length, then $\text{prox}_{th}(x)$ is the *soft-threshold operation*:

$$\text{prox}_{t\ell_1}(x) = \text{sign}(x) \cdot \max\{|x| - t, 0\},$$

where sign and \max are entry-wise, and \cdot denotes the entry-wise product.

The *proximal gradient method* uses the proximal mapping of the nonsmooth part to minimize composite functions:

$$\begin{aligned} x_{k+1} &= x_k - t_k G_{t_k f}(x_k) \\ G_{t_k f}(x_k) &:= \frac{1}{t_k} (x_k - \text{prox}_{t_k h}(x_k - t_k \nabla g(x_k))), \end{aligned}$$

where t_k denotes the k -th step length and $G_{t_k f}(x_k)$ is a *composite gradient step*. Most first-order methods, including SpaRSA and accelerated first-order methods, are variants of this simple method. We note three properties of the composite gradient step:

1. $G_{t_k f}(x_k)$ steps to the minimizer of h plus a simple quadratic model of g near x_k :

$$x_{k+1} = \text{prox}_{t_k h}(x_k - t_k \nabla g(x_k)) \quad (2.2)$$

$$= \arg \min_y t_k h(y) + \frac{1}{2} \|y - x_k + t_k \nabla g(x_k)\|^2 \quad (2.3)$$

$$= \arg \min_y \nabla g(x_k)^T (y - x_k) + \frac{1}{2t_k} \|y - x_k\|^2 + h(y). \quad (2.4)$$

2. $G_{t_k f}(x_k)$ is neither a gradient nor a subgradient of f at any point; rather it is the sum of an explicit gradient and an implicit subgradient:

$$G_{t_k f}(x_k) \in \nabla g(x_k) + \partial h(x_{k+1}).$$

3. $G_{t_k f}(x)$ is zero if and only if x minimizes f .

The third property generalizes the zero gradient optimality condition for smooth functions to composite functions. We shall use the length of $G_f(x)$ to measure the optimality of a point x .

Lemma 2.1. *If ∇g is Lipschitz continuous with constant L_1 , then $\|G_f(x)\|$ satisfies:*

$$\|G_f(x)\| \leq (L_1 + 1) \|x - x^*\|.$$

Proof. The composite gradient steps at x_k and the optimal solution x^* satisfy

$$\begin{aligned} G_f(x_k) &\in \nabla g(x_k) + \partial h(x_k - G_f(x_k)) \\ G_f(x^*) &\in \nabla g(x^*) + \partial h(x^*). \end{aligned}$$

We subtract these two expressions and rearrange to obtain

$$\partial h(x_k - G_f(x_k)) - \partial h(x^*) \ni G_f(x) - (\nabla g(x) - \nabla g(x^*)).$$

∂h is monotone, hence

$$\begin{aligned} 0 &\leq (x - G_f(x) - x^*)^T \partial h(x_k - G_f(x_k)) \\ &= -G_f(x)^T G_f(x) + (x - x^*)^T G_f(x) + G_f(x)^T (\nabla g(x) - \nabla g(x^*)) \\ &\quad - (x - x^*)^T (\nabla g(x) - \nabla g(x^*)). \end{aligned}$$

We drop the last term because it is nonnegative (∇g is monotone) to obtain

$$\begin{aligned} 0 &\leq -\|G_f(x)\|^2 + (x - x^*)G_f(x) + G_f(x)^T(\nabla g(x) - \nabla g(x^*)) \\ &\leq -\|G_f(x)\|^2 - \|G_f(x)\|(\|x - x^*\| + \|\nabla g(x) - \nabla g(x^*)\|). \end{aligned}$$

We rearrange to obtain

$$\|G_f(x)\| \leq \|x - x^*\| + \|\nabla g(x) - \nabla g(x^*)\|.$$

∇g is Lipschitz continuous, hence

$$\|G_f(x)\| \leq (L_1 + 1) \|x - x^*\|.$$

□

2.2 Proximal Newton-type methods

Proximal Newton-type methods use a local quadratic model (in lieu of the simple quadratic model in the proximal gradient method (2.4)) to account for the curvature of g . A local quadratic model of g at x_k is

$$\hat{g}_k(y) = \nabla g(x_k)^T(x - x_k) + \frac{1}{2}(y - x_k)^T H_k(y - x_k),$$

where H_k denotes an approximation to $\nabla^2 g(x_k)$. A proximal Newton-type search direction Δx_k solves the subproblem

$$\Delta x_k = \arg \min_d \hat{f}_k(x_k + d) := \hat{g}_k(x_k + d) + h(x_k + d). \quad (2.5)$$

There are many strategies for choosing H_k . If we choose H_k to be $\nabla^2 g(x_k)$, then we obtain the *proximal Newton method*. If we build an approximation to $\nabla^2 g(x_k)$ using changes measured in ∇g according to a quasi-Newton strategy, we obtain a *proximal quasi-Newton method*. If the problem is large, we can use limited memory quasi-Newton updates to reduce memory usage. Generally speaking, most strategies for choosing Hessian approximations for Newton-type methods (for minimizing smooth functions) can be adapted to choosing H_k in the context of proximal Newton-type methods.

We can also express a proximal Newton-type search direction using *scaled proximal mappings*. This lets us interpret a proximal Newton-type search direction as a “composite Newton step” and reveals a connection with the composite gradient step.

Definition 2.2. *Let h be a convex function and H , a positive definite matrix. Then the scaled proximal mapping of h at x is*

$$\text{prox}_h^H(x) := \arg \min_{y \in \mathbf{R}^n} h(y) + \frac{1}{2} \|y - x\|_H^2. \quad (2.6)$$

Scaled proximal mappings share many properties with (unscaled) proximal mappings:

1. $\text{prox}_h^H(x)$ exists and is unique for $x \in \text{dom } h$ because the proximity function is strongly convex if H is positive definite.
2. Let $\partial h(x)$ be the subdifferential of h at x . Then $\text{prox}_h^H(x)$ satisfies

$$H(x - \text{prox}_h^H(x)) \in \partial h(\text{prox}_h^H(x)). \quad (2.7)$$

3. $\text{prox}_h^H(x)$ is *firmly nonexpansive* in the H -norm. That is, if $u = \text{prox}_h^H(x)$ and $v = \text{prox}_h^H(y)$, then

$$(u - v)^T H(x - y) \geq \|u - v\|_H^2,$$

and the Cauchy-Schwarz inequality implies $\|u - v\|_H \leq \|x - y\|_H$.

We can express a proximal Newton-type search direction as a “composite Newton step” using scaled proximal mappings:

$$\Delta x = \text{prox}_h^H(x - H^{-1}\nabla g(x)) - x. \quad (2.8)$$

We use (2.7) to deduce that a proximal Newton search direction satisfies

$$H(H^{-1}\nabla g(x) - \Delta x) \in \partial h(x + \Delta x).$$

We simplify to obtain

$$H\Delta x \in -\nabla g(x) - \partial h(x + \Delta x). \quad (2.9)$$

Thus a proximal Newton-type search direction, like the composite gradient step, combines an explicit gradient with an implicit subgradient. Note this expression yields the Newton system in the case of smooth functions (*i.e.*, h is zero).

Proposition 2.3 (Search direction properties). If H is positive definite, then Δx in (2.5) satisfies

$$f(x_+) \leq f(x) + t(\nabla g(x)^T \Delta x + h(x + \Delta x) - h(x)) + O(t^2), \quad (2.10)$$

$$\nabla g(x)^T \Delta x + h(x + \Delta x) - h(x) \leq -\Delta x^T H \Delta x. \quad (2.11)$$

Proof. For $t \in (0, 1]$,

$$\begin{aligned} f(x_+) - f(x) &= g(x_+) - g(x) + h(x_+) - h(x) \\ &\leq g(x_+) - g(x) + th(x + \Delta x) + (1 - t)h(x) - h(x) \\ &= g(x_+) - g(x) + t(h(x + \Delta x) - h(x)) \\ &= \nabla g(x)^T (t\Delta x) + t(h(x + \Delta x) - h(x)) + O(t^2), \end{aligned}$$

which proves (2.10).

Since Δx steps to the minimizer of \hat{f} (2.5), $t\Delta x$ satisfies

$$\begin{aligned} \nabla g(x)^T \Delta x + \frac{1}{2} \Delta x^T H \Delta x + h(x + \Delta x) \\ \leq \nabla g(x)^T (t\Delta x) + \frac{1}{2} t^2 \Delta x^T H \Delta x + h(x_+) \\ \leq t \nabla g(x)^T \Delta x + \frac{1}{2} t^2 \Delta x^T H \Delta x + t h(x + \Delta x) + (1-t) h(x). \end{aligned}$$

We rearrange and then simplify:

$$\begin{aligned} (1-t) \nabla g(x)^T \Delta x + \frac{1}{2} (1-t^2) \Delta x^T H \Delta x + (1-t)(h(x + \Delta x) - h(x)) &\leq 0 \\ \nabla g(x)^T \Delta x + \frac{1}{2} (1+t) \Delta x^T H \Delta x + h(x + \Delta x) - h(x) &\leq 0 \\ \nabla g(x)^T \Delta x + h(x + \Delta x) - h(x) &\leq -\frac{1}{2} (1+t) \Delta x^T H \Delta x. \end{aligned}$$

Finally, we let $t \rightarrow 1$ and rearrange to obtain (2.11). \square

Proposition 2.3 implies the search direction is a descent direction for f because we can substitute (2.11) into (2.10) to obtain

$$f(x_+) \leq f(x) - t \Delta x^T H \Delta x + O(t^2). \quad (2.12)$$

Proposition 2.4. Suppose H is positive definite. Then x^* is an optimal solution if and only if at x^* the search direction Δx (2.5) is zero.

Proof. If Δx at x^* is nonzero, it is a descent direction for f at x^* . Hence, x^* cannot be a minimizer of f . If $\Delta x = 0$, then x is the minimizer of \hat{f} . Thus

$$\nabla g(x)^T (td) + \frac{1}{2} t^2 d^T H d + h(x + td) - h(x) \geq 0$$

for all $t > 0$ and d . We rearrange to obtain

$$h(x + td) - h(x) \geq -t \nabla g(x)^T d - \frac{1}{2} t^2 d^T H d. \quad (2.13)$$

Let $Df(x, d)$ be the directional derivative of f at x in the direction d :

$$\begin{aligned} Df(x, d) &= \lim_{t \rightarrow 0} \frac{f(x + td) - f(x)}{t} \\ &= \lim_{t \rightarrow 0} \frac{g(x + td) - g(x) + h(x + td) - h(x)}{t} \\ &= \lim_{t \rightarrow 0} \frac{t \nabla g(x)^T d + O(t^2) + h(x + td) - h(x)}{t}. \end{aligned} \quad (2.14)$$

We substitute (2.13) into (2.14) to obtain

$$\begin{aligned} Df(x, u) &\geq \lim_{t \rightarrow 0} \frac{t \nabla g(x)^T d + O(t^2) - \frac{1}{2} t^2 d^T H d - t \nabla g(x)^T d}{t} \\ &= \lim_{t \rightarrow 0} \frac{-\frac{1}{2} t^2 d^T H d + O(t^2)}{t} = 0. \end{aligned}$$

Since f is convex, x is an optimal solution if and only if $\Delta x = 0$. \square

In a few special cases we can derive a closed form expression for the proximal Newton search direction, but we must usually resort to an iterative method. The user should choose an iterative method that exploits the properties of h . *E.g.*, if h is the ℓ_1 norm, then (block) coordinate descent methods combined with an active set strategy are known to be very efficient for these problems [9].

We use a line search procedure to select a step length t that satisfies a sufficient descent condition:

$$f(x_+) \leq f(x) + \alpha t \Delta \quad (2.15)$$

$$\Delta := \nabla g(x)^T \Delta x + h(x + \Delta x) - h(x), \quad (2.16)$$

where $\alpha \in (0, 0.5)$ can be interpreted as the fraction of the decrease in f predicted by linear extrapolation that we will accept. A simple example of a line search procedure is called *backtracking line search* [4].

Lemma 2.5. *Suppose $H \succeq mI$ for some $m > 0$ and ∇g is Lipschitz continuous with constant L_1 . Then there exists κ such that*

$$t \leq \min \left\{ 1, \frac{2}{\kappa} (1 - \alpha) \right\} \quad (2.17)$$

satisfies the sufficient descent condition (2.16).

Proof. We can bound the decrease at each iteration by

$$\begin{aligned} f(x_+) - f(x) &= g(x_+) - g(x) + h(x_+) - h(x) \\ &\leq \int_0^1 \nabla g(x + s(t\Delta x))^T (t\Delta x) ds + th(x + \Delta x) + (1 - t)h(x) - h(x) \\ &= \nabla g(x)^T (t\Delta x) + t(h(x + \Delta x) - h(x)) \\ &\quad + \int_0^1 (\nabla g(x + s(t\Delta x)) - \nabla g(x))^T (t\Delta x) ds \\ &\leq t (\nabla g(x)^T (t\Delta x) + h(x + \Delta x) - h(x)) \\ &\quad + \int_0^1 \|\nabla g(x + s(\Delta x)) - \nabla g(x)\| \|\Delta x\| ds. \end{aligned}$$

Since ∇g is Lipschitz continuous with constant L_1 ,

$$\begin{aligned} f(x_+) - f(x) &\leq t \left(\nabla g(x)^T \Delta x + h(x + \Delta x) - h(x) + \frac{L_1 t}{2} \|\Delta x\|^2 \right) \\ &= t \left(\Delta + \frac{L_1 t}{2} \|\Delta x\|^2 \right), \end{aligned} \quad (2.18)$$

where we use (2.11). If we choose $t \leq \frac{2}{\kappa} (1 - \alpha)$, $\kappa = L_1/m$, then

$$\begin{aligned} \frac{L_1 t}{2} \|\Delta x\|^2 &\leq m(1 - \alpha) \|\Delta x\|^2 \\ &\leq (1 - \alpha) \Delta x^T H \Delta x \\ &\leq -(1 - \alpha) \Delta, \end{aligned} \quad (2.19)$$

where we again use (2.11). We substitute (2.19) into (2.18) to obtain

$$f(x_+) - f(x) \leq t(\Delta - (1 - \alpha)\Delta) = t(\alpha\Delta).$$

□

Algorithm 1 A generic proximal Newton-type method

Require: starting point $x_0 \in \text{dom } f$

- 1: **repeat**
 - 2: Choose H_k , a positive definite approximation to the Hessian.
 - 3: Solve the subproblem for a search direction:
 $\Delta x_k \leftarrow \arg \min_d \nabla g(x_k)^T d + \frac{1}{2} d^T H_k d + h(x_k + d).$
 - 4: Select t_k with a backtracking line search.
 - 5: Update: $x_{k+1} \leftarrow x_k + t_k \Delta x_k.$
 - 6: **until** stopping conditions are satisfied.
-

2.3 Inexact proximal Newton-type methods

Inexact proximal Newton-type methods solve subproblem (2.5) approximately to obtain inexact search directions. These methods can be more efficient than their exact counterparts because they require less computational expense per iteration. In fact, many practical implementations of proximal Newton-type methods such as `glmnet`, `LIBLINEAR`, and `QUIC` use inexact search directions.

In practice, how exactly (or inexactly) we solve the subproblem is critical to the efficiency and reliability of the method. The practical implementations of proximal Newton-type methods we mentioned use a variety of heuristics to decide how accurately to solve the subproblem. Although these methods perform admirably in practice, there are few results on how inexact solutions to the subproblem affect their convergence behavior.

First we propose an adaptive stopping condition for the subproblem. Then in section 3 we analyze the convergence behavior of inexact Newton-type methods. Finally, in section 4 we conduct computational experiments to compare the performance of our stopping condition against commonly used heuristics.

Our stopping condition is motivated by the adaptive stopping condition used by *inexact Newton-type methods* for minimizing smooth functions:

$$\|\nabla \hat{g}_k(x_k + \Delta x_k)\| \leq \eta_k \|\nabla g(x_k)\|, \quad (2.20)$$

where η_k is called a *forcing term* because it forces the left-hand side to be small. We generalize (2.20) to composite functions by substituting composite gradients into (2.20) and scaling the norm:

$$\|\nabla \hat{g}_k(x_k) + \partial h(x_k + \Delta x_k)\|_{H_k^{-1}} \leq \eta_k \|G_f(x_k)\|_{H_k^{-1}}. \quad (2.21)$$

Following Eisenstat and Walker [8], we set η_k based on how well \hat{g}_{k-1} approximates g near x_k :

$$\eta_k = \min \left\{ 0.1, \frac{\|\nabla \hat{g}_{k-1}(x_k) - \nabla g(x_k)\|}{\|\nabla g(x_{k-1})\|} \right\}. \quad (2.22)$$

This choice yields desirable convergence results and performs admirably in practice.

Intuitively, we should solve the subproblem exactly if (i) x_k is close to the optimal solution, and (ii) \hat{f}_k is a good model of f near x_k . If (i), then we seek to preserve the fast local convergence behavior of proximal Newton-type methods; if (ii), then minimizing \hat{f}_k is a good surrogate for minimizing f . In these cases, (2.21) and (2.22) ensure the subproblem is solved accurately.

We can derive an expression like (2.9) for an inexact search direction in terms of an explicit gradient, an implicit subgradient, and a residual term r_k . This reveals connections to the inexact Newton search direction in the case of smooth problems. (2.21) is equivalent to

$$0 \in \nabla \hat{g}_k(x_k) + \partial h(x_k + \Delta x_k) + r_k,$$

for some r_k such that $\|r_k\|_{\nabla^2 g(x_k)^{-1}} \leq \eta_k \|G_f(x_k)\|_{H_k^{-1}}$. Hence an inexact search direction satisfies

$$H \Delta x_k \in -\nabla g(x_k) - \partial h(x_k + \Delta x_k) + r_k. \quad (2.23)$$

3 Convergence results

Our first result guarantees proximal Newton-type methods converge globally to some optimal solution x^* . We assume $\{H_k\}$ are sufficiently positive definite; *i.e.*, $H_k \succeq mI$ for some $m > 0$. This assumption is required to guarantee the methods are executable, *i.e.* there exist step lengths that satisfy the sufficient descent condition (*cf.* Lemma 2.5).

Theorem 3.1. *If $H_k \succeq mI$ for some $m > 0$, then x_k converges to an optimal solution starting at any $x_0 \in \text{dom } f$.*

Proof. $f(x_k)$ is decreasing because Δx_k is always a descent direction (2.12) and there exist step lengths satisfying the sufficient descent condition (2.16) (*cf.* Lemma 2.5):

$$f(x_k) - f(x_{k+1}) \leq \alpha t_k \Delta_k \leq 0.$$

$f(x_k)$ must converge to some limit (we assumed f is closed and the optimal value is attained); hence $t_k \Delta_k$ must decay to zero. t_k is bounded away from zero because sufficiently small step lengths attain sufficient descent; hence Δ_k must decay to zero. We use (2.11) to deduce that Δx_k also converges to zero:

$$\|\Delta x_k\|^2 \leq \frac{1}{m} \Delta x_k^T H_k \Delta x_k \leq -\frac{1}{m} \Delta_k.$$

Δx_k is zero if and only if x is an optimal solution (*cf.* Proposition 2.4), hence x_k converges to some x^* . \square

3.1 Convergence of the proximal Newton method

The proximal Newton method uses the exact Hessian of the smooth part g in the second-order model of f , *i.e.* $H_k = \nabla^2 g(x_k)$. This method converges q -quadratically:

$$\|x_{k+1} - x^*\| = O(\|x_k - x^*\|^2),$$

subject to standard assumptions on the smooth part: we require g to be locally strongly convex and $\nabla^2 g$ to be locally Lipschitz continuous, *i.e.* g is strongly convex and Lipschitz continuous in a ball around x^* . These are standard assumptions for proving that Newton's method for minimizing smooth functions converges q -quadratically.

First, we prove an auxiliary result: step lengths of unity satisfy the sufficient descent condition after sufficiently many iterations.

Lemma 3.2. *Suppose (i) g is locally strongly convex with constant m and (ii) $\nabla^2 g$ is locally Lipschitz continuous with constant L_2 . If we choose $H_k = \nabla^2 g(x_k)$, then the unit step length satisfies the sufficient decrease condition (2.16) for k sufficiently large.*

Proof. Since $\nabla^2 g$ is locally Lipschitz continuous with constant L_2 ,

$$g(x + \Delta x) \leq g(x) + \nabla g(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 g(x) \Delta x + \frac{L_2}{6} \|\Delta x\|^3.$$

We add $h(x + \Delta x)$ to both sides to obtain

$$\begin{aligned} f(x + \Delta x) &\leq g(x) + \nabla g(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 g(x) \Delta x \\ &\quad + \frac{L_2}{6} \|\Delta x\|^3 + h(x + \Delta x). \end{aligned}$$

We then add and subtract $h(x)$ from the right-hand side to obtain

$$\begin{aligned} f(x + \Delta x) &\leq g(x) + h(x) + \nabla g(x)^T \Delta x + h(x + \Delta x) - h(x) \\ &\quad + \frac{1}{2} \Delta x^T \nabla^2 g(x) \Delta x + \frac{L_2}{6} \|\Delta x\|^3 \\ &\leq f(x) + \Delta + \frac{1}{2} \Delta x^T \nabla^2 g(x) \Delta x + \frac{L_2}{6} \|\Delta x\|^3 \\ &\leq f(x) + \Delta - \frac{1}{2} \Delta + \frac{L_2}{6m} \|\Delta x\| \Delta, \end{aligned}$$

where we use (2.11) and (2.16). We rearrange to obtain

$$\begin{aligned} f(x + \Delta x) - f(x) &\leq \frac{1}{2} \Delta + \frac{1}{2} \Delta x^T \nabla^2 g(x) \Delta x - \frac{L_2}{6m} \Delta \|\Delta x\| \\ &\leq \left(\frac{1}{2} - \frac{L_2}{6m} \right) \Delta + o(\|\Delta x\|^2). \end{aligned}$$

We can show Δx_k decays to zero via the same argument that we used to prove Theorem 3.1. Hence, if k is sufficiently large, $f(x_k + \Delta x_k) - f(x_k) < \frac{1}{2} \Delta_k$. \square

Theorem 3.3. Suppose (i) g is locally strongly convex with constant m , and (ii) $\nabla^2 g$ is locally Lipschitz continuous with constant L_2 . Then the proximal Newton method converges q -quadratically to x^* .

Proof. The assumptions of Lemma 3.2 are satisfied; hence unit step lengths satisfy the sufficient descent condition after sufficiently many steps:

$$x_{k+1} = x_k + \Delta x_k = \text{prox}_h^{\nabla^2 g(x_k)}(x_k - \nabla^2 g(x_k)^{-1} \nabla g(x_k)).$$

$\text{prox}_h^{\nabla^2 g(x_k)}$ is firmly nonexpansive in the $\nabla^2 g(x_k)$ -norm, hence

$$\begin{aligned} & \|x_{k+1} - x^*\|_{\nabla^2 g(x_k)} \\ &= \left\| \text{prox}_h^{\nabla^2 g(x_k)}(x_k - \nabla^2 g(x_k)^{-1} \nabla g(x_k)) \right. \\ &\quad \left. - \text{prox}_h^{\nabla^2 g(x_k)}(x^* - \nabla^2 g(x_k)^{-1} \nabla g(x^*)) \right\|_{\nabla^2 g(x_k)} \\ &\leq \left\| x_k - x^* + \nabla^2 g(x_k)^{-1} (\nabla g(x^*) - \nabla g(x_k)) \right\|_{\nabla^2 g(x_k)} \\ &\leq \frac{1}{\sqrt{m}} \left\| \nabla^2 g(x_k)(x_k - x^*) - \nabla g(x_k) + \nabla g(x^*) \right\|. \end{aligned}$$

$\nabla^2 g$ is locally Lipschitz continuous with constant L_2 ; hence

$$\left\| \nabla^2 g(x_k)(x_k - x^*) - \nabla g(x_k) + \nabla g(x^*) \right\| \leq \frac{L_2}{2} \|x_k - x^*\|^2.$$

We deduce that x_k converges to x^* quadratically:

$$\|x_{k+1} - x^*\| \leq \frac{1}{\sqrt{m}} \|x_{k+1} - x^*\|_{\nabla^2 g(x_k)} \leq \frac{L_2}{2m} \|x_k - x^*\|^2.$$

□

3.2 Convergence of proximal quasi-Newton methods

If the sequence $\{H_k\}$ satisfy the Dennis-Moré criterion [7], namely

$$\frac{\|(H_k - \nabla^2 g(x^*)) (x_{k+1} - x_k)\|}{\|x_{k+1} - x_k\|} \rightarrow 0, \quad (3.1)$$

then we can prove that a proximal quasi-Newton method converges q -superlinearly:

$$\|x_{k+1} - x^*\| \leq o(\|x_k - x^*\|).$$

We also require g to be locally strongly convex and $\nabla^2 g$ to be locally Lipschitz continuous. These are the same assumptions required to prove quasi-Newton methods for minimizing smooth functions converge superlinearly.

First, we prove two auxiliary results: (i) step lengths of unity satisfy the sufficient descent condition after sufficiently many iterations, and (ii) the proximal quasi-Newton step is close to the proximal Newton step.

Lemma 3.4. *Suppose g is twice continuously differentiable and $\nabla^2 g$ is locally Lipschitz continuous with constant L_2 . If $\{H_k\}$ satisfy the Dennis-Moré criterion and their eigenvalues are bounded, then the unit step length satisfies the sufficient descent condition (2.16) after sufficiently many iterations.*

Proof. The proof is very similar to the proof of Lemma 3.2, and we defer the details to Appendix A. \square

The proof of the next result mimics the analysis of Tseng and Yun [25].

Proposition 3.5. Suppose H and \hat{H} are positive definite matrices with bounded eigenvalues: $mI \preceq H \preceq MI$ and $\hat{m}I \preceq \hat{H} \preceq \hat{M}I$. Let Δx and $\Delta \hat{x}$ be the search directions generated using H and \hat{H} respectively:

$$\begin{aligned}\Delta x &= \text{prox}_h^H(x - H^{-1}\nabla g(x)) - x, \\ \Delta \hat{x} &= \text{prox}_h^{\hat{H}}(x - \hat{H}^{-1}\nabla g(x)) - x.\end{aligned}$$

Then there exists $\bar{\theta}$ such that these two search directions satisfy

$$\|\Delta x - \Delta \hat{x}\| \leq \sqrt{\frac{1+\bar{\theta}}{m}} \|(\hat{H} - H)\Delta x\|^{1/2} \|\Delta x\|^{1/2}.$$

Proof. By (2.5) and Fermat's rule, Δx and $\Delta \hat{x}$ are also the solutions to

$$\begin{aligned}\Delta x &= \arg \min_d \nabla g(x)^T d + \Delta x^T H d + h(x + d), \\ \Delta \hat{x} &= \arg \min_d \nabla g(x)^T d + \Delta \hat{x}^T \hat{H} d + h(x + d).\end{aligned}$$

Hence Δx and $\Delta \hat{x}$ satisfy

$$\begin{aligned}\nabla g(x)^T \Delta x + \Delta x^T H \Delta x + h(x + \Delta x) \\ \leq \nabla g(x)^T \Delta \hat{x} + \Delta \hat{x}^T H \Delta \hat{x} + h(x + \Delta \hat{x})\end{aligned}$$

and

$$\begin{aligned}\nabla g(x)^T \Delta \hat{x} + \Delta \hat{x}^T \hat{H} \Delta \hat{x} + h(x + \Delta \hat{x}) \\ \leq \nabla g(x)^T \Delta x + \Delta x^T \hat{H} \Delta x + h(x + \Delta x).\end{aligned}$$

We sum these two inequalities and rearrange to obtain

$$\Delta x^T H \Delta x - \Delta x^T (H + \hat{H}) \Delta \hat{x} + \Delta \hat{x}^T \hat{H} \Delta \hat{x} \leq 0.$$

We then complete the square on the left side and rearrange to obtain

$$\begin{aligned}\Delta x^T H \Delta x - 2\Delta x^T H \Delta \hat{x} + \Delta \hat{x}^T H \Delta \hat{x} \\ \leq \Delta x^T (\hat{H} - H) \Delta \hat{x} + \Delta \hat{x}^T (H - \hat{H}) \Delta \hat{x}.\end{aligned}$$

The left side is $\|\Delta x - \Delta \hat{x}\|_H^2$ and the eigenvalues of H are bounded. Thus

$$\begin{aligned}\|\Delta x - \Delta \hat{x}\| &\leq \frac{1}{\sqrt{m}} \left(\Delta x^T (\hat{H} - H) \Delta x + \Delta \hat{x}^T (H - \hat{H}) \Delta \hat{x} \right)^{1/2} \\ &\leq \frac{1}{\sqrt{m}} \|(\hat{H} - H) \Delta \hat{x}\|^{1/2} (\|\Delta x\| + \|\Delta \hat{x}\|)^{1/2}.\end{aligned}\quad (3.2)$$

We use a result due to Tseng and Yun (cf. Lemma 3 in [25]) to bound the term $(\|\Delta x\| + \|\Delta \hat{x}\|)$. Let P denote $\hat{H}^{-1/2} H \hat{H}^{-1/2}$. Then $\|\Delta x\|$ and $\|\Delta \hat{x}\|$ satisfy

$$\|\Delta x\| \leq \left(\frac{\hat{M} \left(1 + \lambda_{\max}(P) + \sqrt{1 - 2\lambda_{\min}(P) + \lambda_{\max}(P)^2} \right)}{2m} \right) \|\Delta \hat{x}\|.$$

We denote the constant in parentheses by $\bar{\theta}$ and conclude that

$$\|\Delta x\| + \|\Delta \hat{x}\| \leq (1 + \bar{\theta}) \|\Delta \hat{x}\|. \quad (3.3)$$

We substitute (3.3) into (3.2) to obtain

$$\|\Delta x - \Delta \hat{x}\|^2 \leq \sqrt{\frac{1 + \bar{\theta}}{m}} \|(\hat{H} - H) \Delta \hat{x}\|^{1/2} \|\Delta \hat{x}\|^{1/2}.$$

□

Theorem 3.6. *Suppose (i) g is twice continuously differentiable and locally strongly convex, (ii) $\nabla^2 g$ is locally Lipschitz continuous with constant L_2 . If $\{H_k\}$ satisfy the Dennis-Moré criterion and their eigenvalues are bounded, then a proximal quasi-Newton method converges q -superlinearly to x^* .*

Proof. The assumptions of Lemma 3.4 are satisfied; hence unit step lengths satisfy the sufficient descent condition after sufficiently many iterations:

$$x_{k+1} = x_k + \Delta x_k.$$

Since the proximal Newton method converges q -quadratically (cf. Theorem 3.3),

$$\begin{aligned}\|x_{k+1} - x^*\| &\leq \|x_k + \Delta x_k^{\text{nt}} - x^*\| + \|\Delta x_k - \Delta x_k^{\text{nt}}\| \\ &\leq \frac{L_2}{m} \|x_k^{\text{nt}} - x^*\|^2 + \|\Delta x_k - \Delta x_k^{\text{nt}}\|,\end{aligned}\quad (3.4)$$

where Δx_k^{nt} denotes the proximal-Newton search direction. We use Proposition 3.5 to bound the second term:

$$\|\Delta x_k - \Delta x_k^{\text{nt}}\| \leq \sqrt{\frac{1 + \bar{\theta}}{m}} \|(\nabla^2 g(x_k) - H_k) \Delta x_k\|^{1/2} \|\Delta x_k\|^{1/2}. \quad (3.5)$$

$\nabla^2 g$ is Lipschitz continuous and Δx_k satisfies the Dennis-Moré criterion; hence

$$\begin{aligned}\|(\nabla^2 g(x_k) - H_k) \Delta x_k\| &\leq \|(\nabla^2 g(x_k) - \nabla^2 g(x^*)) \Delta x_k\| \\ &\quad + \|(\nabla^2 g(x^*) - H_k) \Delta x_k\| \\ &\leq L_2 \|x_k - x^*\| \|\Delta x_k\| + o(\|\Delta x_k\|).\end{aligned}$$

$\|\Delta x_k\|$ is within some constant $\bar{\theta}_k$ of $\|\Delta x_k^{\text{nt}}\|$ (cf. Lemma 3 in [25]), and we know the proximal Newton method converges q -quadratically. Thus

$$\begin{aligned}\|\Delta x_k\| &\leq \bar{\theta}_k \|\Delta x_k^{\text{nt}}\| = \bar{\theta}_k \|x_{k+1}^{\text{nt}} - x_k\| \\ &\leq \bar{\theta}_k (\|x_{k+1}^{\text{nt}} - x^*\| + \|x_k - x^*\|) \\ &\leq O(\|x_k - x^*\|^2) + \bar{\theta}_k \|x_k - x^*\|.\end{aligned}$$

We substitute these expressions into (3.5) to obtain

$$\|\Delta x_k - \Delta x_k^{\text{nt}}\| = o(\|x_k - x^*\|).$$

We substitute this expression into (3.4) to obtain

$$\|x_{k+1} - x^*\| \leq \frac{L_2}{m} \|x_k^{\text{nt}} - x^*\|^2 + o(\|x_k - x^*\|),$$

and we deduce that x_k converges to x^* superlinearly. \square

3.3 Convergence of the inexact proximal Newton method

We make the same assumptions made by Dembo et al. in their analysis of *inexact Newton methods* for minimizing smooth functions [6]: (i) x_k is close to x^* and (ii) the unit step length is eventually accepted. We prove the inexact proximal Newton method (i) converges q -linearly if the forcing terms η_k are smaller than some $\bar{\eta}$, and (ii) converges q -superlinearly if the forcing terms decay to zero.

First, we prove a consequence of the smoothness of g . Then, we use this result to prove the inexact proximal Newton method converges locally subject to standard assumptions on g and η_k .

Lemma 3.7. *Suppose g is locally strongly convex and $\nabla^2 g$ is locally Lipschitz continuous. If x_k sufficiently close to x^* , then for any x ,*

$$\|x - x^*\|_{\nabla^2 g(x^*)} \leq (1 + \epsilon) \|x - x^*\|_{\nabla^2 g(x_k)}.$$

Proof. We first expand $\nabla^2 g(x^*)^{1/2}(x - x^*)$ to obtain

$$\begin{aligned}\nabla^2 g(x^*)^{1/2}(x - x^*) &= \left(\nabla^2 g(x^*)^{1/2} - \nabla^2 g(x_k)^{1/2} \right) (x - x^*) + \nabla^2 g(x_k)^{1/2}(x - x^*) \\ &= \left(\nabla^2 g(x^*)^{1/2} - \nabla^2 g(x_k)^{1/2} \right) \nabla^2 g(x_k)^{-1/2} \nabla^2 g(x_k)^{1/2}(x - x^*) \\ &\quad + \nabla^2 g(x_k)^{1/2}(x - x^*) \\ &= \left(I + \left(\nabla^2 g(x^*)^{1/2} - \nabla^2 g(x_k)^{1/2} \right) \nabla^2 g(x_k)^{-1/2} \right) \nabla^2 g(x_k)^{1/2}(x - x^*).\end{aligned}$$

We take norms to obtain

$$\begin{aligned}\|x - x^*\|_{\nabla^2 g(x^*)} &\leq \|I + (\nabla^2 g(x^*)^{1/2} - \nabla^2 g(x_k)^{1/2}) \nabla^2 g(x_k)^{-1/2}\| \|x - x^*\|_{\nabla^2 g(x_k)}.\end{aligned}\tag{3.6}$$

If g is locally strongly convex with constant m and x_k is sufficiently close to x^* , then

$$\|\nabla^2 g(x^*)^{1/2} - \nabla^2 g(x_k)^{1/2}\| \leq \sqrt{m}\epsilon.$$

We substitute this bound into (3.6) to deduce that

$$\|x - x^*\|_{\nabla^2 g(x^*)} \leq (1 + \epsilon) \|x - x^*\|_{\nabla^2 g(x_k)}.$$

□

Theorem 3.8. Suppose (i) g is locally strongly convex with constant m , (ii) $\nabla^2 g$ is locally Lipschitz continuous with constant L_2 , and (iii) there exists L_G such that the composite gradient step satisfies

$$\|G_f(x_k)\|_{\nabla^2 g(x_k)^{-1}} \leq L_G \|x_k - x^*\|_{\nabla^2 g(x_k)}. \quad (3.7)$$

1. If η_k is smaller than some $\bar{\eta} < \frac{1}{L_G}$, then an inexact proximal Newton method converges q -linearly to x^* .
2. If η_k decays to zero, then an inexact proximal Newton method converges q -superlinearly to x^* .

Proof. We use Lemma 3.7 to deduce

$$\|x_{k+1} - x^*\|_{\nabla^2 g(x^*)} \leq (1 + \epsilon_1) \|x_k - x^*\|_{\nabla^2 g(x_k)}. \quad (3.8)$$

We use the monotonicity of ∂h to bound $\|x_k - x^*\|_{\nabla^2 g(x_k)}$. First, Δx_k satisfies

$$\nabla^2 g(x_k)(x_{k+1} - x^*) \in -\nabla g(x_k) - \partial h(x_{k+1}) + r_k \quad (3.9)$$

(cf. (2.23)). Also, the exact proximal Newton step at x^* (trivially) satisfies

$$\nabla^2 g(x_k)(x^* - x^*) \in -\nabla g(x^*) - \partial h(x^*). \quad (3.10)$$

Subtracting (3.10) from (3.9) and rearranging, we obtain

$$\begin{aligned} & \partial h(x_{k+1}) - \partial h(x^*) \\ & \in \nabla^2 g(x_k)(x_k - x_{k+1} - x^* + x^*) - \nabla g(x_k) + \nabla g(x^*) + r_k. \end{aligned}$$

Since ∂h is monotone,

$$\begin{aligned} 0 & \leq (x_{k+1} - x^*)^T (\partial h(x_{k+1}) - \partial h(x^*)) \\ & = (x_{k+1} - x^*)^T \nabla^2 g(x_k)(x^* - x_{k+1}) + (x_{k+1} - x^*)^T (\nabla^2 g(x_k)(x_k - x^*) \\ & \quad - \nabla g(x_k) + \nabla g(x^*) + r_k) \\ & = (x_{k+1} - x^*)^T \nabla^2 g(x_k) (x_k - x^* + \nabla^2 g(x_k)^{-1} (\nabla g(x^*) - \nabla g(x_k) + r_k)) \\ & \quad - \|x_{k+1} - x^*\|_{\nabla^2 g(x_k)}. \end{aligned}$$

We taking norms to obtain

$$\begin{aligned} \|x_{k+1} - x^*\|_{\nabla^2 g(x_k)} &\leq \|x_k - x^* + \nabla^2 g(x_k)^{-1} (\nabla g(x^*) - \nabla g(x_k))\|_{\nabla^2 g(x_k)} \\ &\quad + \eta_k \|r_k\|_{\nabla^2 g(x_k)^{-1}}. \end{aligned}$$

If x_k is sufficiently close to x^* , for any $\epsilon_2 > 0$ we have

$$\|x_k - x^* + \nabla^2 g(x_k)^{-1} (\nabla g(x^*) - \nabla g(x_k))\|_{\nabla^2 g(x_k)} \leq \epsilon_2 \|x_k - x^*\|_{\nabla g^2(x^*)}$$

and hence

$$\|x_{k+1} - x^*\|_{\nabla^2 g(x^*)} \leq \epsilon_2 \|x_k - x^*\|_{\nabla g^2(x^*)} + \eta_k \|r_k\|_{\nabla^2 g(x_k)^{-1}}. \quad (3.11)$$

We substitute (3.11) into (3.8) to obtain

$$\|x_{k+1} - x^*\|_{\nabla^2 g(x^*)} \leq (1 + \epsilon_1)(\epsilon_2 \|x_k - x^*\|_{\nabla g^2(x^*)} + \eta_k \|r_k\|_{\nabla^2 g(x_k)^{-1}}). \quad (3.12)$$

Since Δx_k satisfies the adaptive stopping condition (2.21),

$$\|r_k\|_{\nabla^2 g(x_k)^{-1}} \leq \eta_k \|G_f(x_k)\|_{\nabla^2 g(x_k)^{-1}},$$

and since there exists L_G such that G_f satisfies (3.7),

$$\|r_k\|_{\nabla^2 g(x_k)^{-1}} \leq \eta_k L_G \|x_k - x^*\|_{\nabla g^2(x^*)}. \quad (3.13)$$

We substitute (3.13) into (3.12) to obtain

$$\|x_{k+1} - x^*\|_{\nabla^2 g(x^*)} \leq (1 + \epsilon_1)(\epsilon_2 + \eta_k L_G) \|x_k - x^*\|_{\nabla g^2(x^*)}.$$

If η_k is smaller than some

$$\bar{\eta} < \frac{1}{L_G} \left(\frac{1}{(1 + \epsilon_1)} - \epsilon_2 \right) < \frac{1}{L_G}$$

then x_k converges q -linearly to x^* . If η_k decays to zero (the smoothness of g lets ϵ_1, ϵ_2 decay to zero), then x_k converges q -superlinearly to x^* . \square

If we assume g is twice continuously differentiable, we can derive an expression for L_G . Combining this result with Theorem 3.8 we deduce the convergence of an inexact Newton method with our adaptive stopping condition (2.21).

Lemma 3.9. *Suppose g is locally strongly convex with constant m , and $\nabla^2 g$ is locally Lipschitz continuous. If x_k is sufficiently close to x^* , there exists κ such that*

$$\|G_f(x_k)\|_{\nabla^2 g(x_k)^{-1}} \leq \left(\sqrt{\kappa}(1 + \epsilon) + \frac{1}{m} \right) \|x_k - x^*\|_{\nabla^2 g(x_k)}.$$

Proof. Since $G_f(x^*)$ is zero,

$$\begin{aligned}
& \|G_f(x_k)\|_{\nabla^2 g(x_k)^{-1}} \\
& \leq \frac{1}{\sqrt{m}} \|G_f(x_k) - G_f(x^*)\| \\
& \leq \frac{1}{\sqrt{m}} \|\nabla g(x_k) - \nabla g(x^*)\| + \frac{1}{\sqrt{m}} \|x_k - x^*\| \\
& \leq \sqrt{\kappa} \|\nabla g(x_k) - \nabla g(x^*)\|_{\nabla^2 g(x_k)^{-1}} + \frac{1}{m} \|x_k - x^*\|_{\nabla^2 g(x_k)}, \quad (3.14)
\end{aligned}$$

where $\kappa = L_2/m$. The second inequality follows from Lemma 2.1. We split $\|\nabla g(x_k) - \nabla g(x^*)\|_{\nabla^2 g(x_k)^{-1}}$ into two terms:

$$\begin{aligned}
& \|\nabla g(x_k) - \nabla g(x^*)\|_{\nabla^2 g(x_k)^{-1}} \\
& = \|\nabla g(x_k) - \nabla g(x^*) + \nabla^2 g(x_k)(x^* - x_k)\|_{\nabla^2 g(x_k)} + \|x_k - x^*\|_{\nabla^2 g(x_k)}.
\end{aligned}$$

If x_k is sufficiently close to x^* , for any $\epsilon_1 > 0$ we have

$$\|\nabla g(x_k) - \nabla g(x^*) + \nabla^2 g(x_k)(x^* - x_k)\|_{\nabla^2 g(x_k)^{-1}} \leq \epsilon_1 \|x_k - x^*\|_{\nabla^2 g(x_k)}.$$

Hence

$$\|\nabla g(x_k) - \nabla g(x^*)\|_{\nabla^2 g(x_k)^{-1}} \leq (1 + \epsilon_1) \|x_k - x^*\|_{\nabla^2 g(x_k)}.$$

We substituting this bound into (3.14) to obtain

$$\|G_f(x_k)\|_{\nabla^2 g(x_k)^{-1}} \leq \left(\sqrt{\kappa}(1 + \epsilon_1) + \frac{1}{m} \right) \|x_k - x^*\|_{\nabla^2 g(x_k)}.$$

We use Lemma 3.7 to deduce

$$\begin{aligned}
\|G_f(x_k)\|_{\nabla^2 g(x_k)^{-1}} & \leq (1 + \epsilon_2) \left(\sqrt{\kappa}(1 + \epsilon_1) + \frac{1}{m} \right) \|x_k - x^*\|_{\nabla^2 g(x_k)} \\
& \leq \left(\sqrt{\kappa}(1 + \epsilon) + \frac{1}{m} \right) \|x_k - x^*\|_{\nabla^2 g(x_k)}.
\end{aligned}$$

□

Corollary 3.10. Suppose (i) g is locally strongly convex with constant m , and (ii) $\nabla^2 g$ is locally Lipschitz continuous with constant L_2 .

1. If η_k is smaller than some $\bar{\eta} < \frac{1}{\sqrt{\kappa}+1/m}$, an inexact proximal Newton method converges q -linearly to x^* .
2. If η_k decays to zero, an inexact proximal Newton method converges q -superlinearly to x^* .

Remark 1. In many cases, we can obtain tighter bounds on $\|G_f(x_k)\|_{\nabla^2 g(x_k)^{-1}}$. E.g., when minimizing smooth functions (h is zero), we can show

$$\|G_f(x_k)\|_{\nabla^2 g(x_k)^{-1}} = \|\nabla g(x_k)\|_{\nabla^2 g(x_k)^{-1}} \leq (1 + \epsilon) \|x_k - x^*\|_{\nabla^2 g(x_k)}.$$

This yields the classic result of Dembo et al.: if η_k is uniformly smaller than one, then the inexact Newton method converges q -linearly.

Finally, we justify our choice of forcing terms: if we choose η_k according to (2.22), then the inexact proximal Newton method converges q -superlinearly.

Theorem 3.11. *Suppose (i) x_0 is sufficiently close to x^* , and (ii) the assumptions of Theorem 3.3 are satisfied. If we choose η_k according to (2.22), then the inexact proximal Newton method converges q -superlinearly.*

Proof. Since the assumptions of Theorem 3.3 are satisfied, x_k converges locally to x^* . Also, since $\nabla^2 g$ is Lipschitz continuous,

$$\begin{aligned} & \|\nabla g(x_k) - \nabla g(x_{k-1}) - \nabla^2 g(x_{k-1}) \Delta x_{k-1}\| \\ & \leq \left(\int_0^1 \|\nabla^2 g(x_{k-1} + s \Delta x_{k-1}) - \nabla^2 g(x^*)\| ds \right) \|\Delta x_{k-1}\| \\ & \quad + \|\nabla^2 g(x^*) - \nabla^2 g(x_{k-1})\| \|\Delta x_{k-1}\| \\ & \leq \left(\int_0^1 L_2 \|x_{k-1} + s \Delta x_{k-1} - x^*\| ds \right) \|\Delta x_{k-1}\| \\ & \quad + L_2 \|x_{k-1} - x^*\| \|\Delta x_{k-1}\|. \end{aligned}$$

We integrate the first term to obtain

$$\int_0^1 L_2 \|x_{k-1} + s \Delta x_{k-1} - x^*\| ds = L_2 \|x_{k-1} - x^*\| + \frac{L_2}{2} \|\Delta x_{k-1}\|.$$

We substituting these expressions into (2.22) to obtain

$$\eta_k \leq L_2 \left(2 \|x_{k-1} - x^*\| + \frac{1}{2} \|\Delta x_{k-1}\| \right) \frac{\|\Delta x_{k-1}\|}{\|\nabla g(x_{k-1})\|}. \quad (3.15)$$

If $\nabla g(x^*) \neq 0$, $\|\nabla g(x)\|$ is bounded away from zero in a neighborhood of x^* . Hence η_k decays to zero and x_k converges q -superlinearly to x^* . Otherwise,

$$\|\nabla g(x_{k-1})\| = \|\nabla g(x_{k-1}) - \nabla g(x^*)\| \geq m \|x_{k-1} - x^*\|. \quad (3.16)$$

We substitute (3.15) and (3.16) into (2.22) to obtain

$$\eta_k \leq \frac{L_2}{m} \left(2 \|x_{k-1} - x^*\| + \frac{\|\Delta x_{k-1}\|}{2} \right) \frac{\|\Delta x_{k-1}\|}{\|x_{k-1} - x^*\|}. \quad (3.17)$$

The triangle inequality yields

$$\|\Delta x_{k-1}\| \leq \|x_k - x^*\| + \|x_{k-1} - x^*\|.$$

We divide by $\|x_{k-1} - x^*\|$ to obtain

$$\frac{\|\Delta x_{k-1}\|}{\|x_{k-1} - x^*\|} \leq 1 + \frac{\|x_k - x^*\|}{\|x_{k-1} - x^*\|}.$$

If k is sufficiently large, x_k converges q -linearly to x^* and hence

$$\frac{\|\Delta x_{k-1}\|}{\|x_{k-1} - x^*\|} \leq 2.$$

We substitute this expression into (3.17) to obtain

$$\eta_k \leq \frac{L_2}{m} (4 \|x_{k-1} - x^*\| + \|\Delta x_{k-1}\|).$$

Hence η_k decays to zero, and x_k converges q -superlinearly to x^* . \square

4 Computational experiments

First we explore how inexact search directions affect the convergence behavior of proximal Newton-type methods on a problem in bioinformatics. We show that choosing the forcing terms according to (2.22) avoids “oversolving” the subproblem. Then we demonstrate the performance of proximal Newton-type methods using a problem in statistical learning. We show that the methods are suited to problems with expensive smooth function evaluations.

4.1 Inverse covariance estimation

Suppose we are given *i.i.d.* samples $x^{(1)}, \dots, x^{(n)}$ from a Gaussian Markov random field (MRF) with unknown inverse covariance matrix $\bar{\Theta}$:

$$\Pr(x; \bar{\Theta}) \propto \exp(x^T \bar{\Theta} x / 2 - \log \det(\bar{\Theta})).$$

We seek a sparse maximum likelihood estimate of the inverse covariance matrix:

$$\hat{\Theta} := \arg \min_{\Theta \in \mathbf{R}^{n \times n}} \text{tr}(\hat{\Sigma} \Theta) - \log \det(\Theta) + \lambda \|\text{vec}(\Theta)\|_1, \quad (4.1)$$

where $\hat{\Sigma}$ denotes the sample covariance matrix. We regularize using an entry-wise ℓ_1 norm to avoid overfitting the data and promote sparse estimates. λ is a parameter that balances goodness-of-fit and sparsity.

We use two datasets: (i) Estrogen, a gene expression dataset consisting of 682 probe sets collected from 158 patients, and (ii) Leukemia, another gene expression dataset consisting of 1255 genes from 72 patients.¹ The features of Estrogen were converted to log-scale and normalized to have zero mean and unit variance. λ was chosen to match the values used in [19].

¹These datasets are available from <http://www.math.nus.edu.sg/~mattohkc/> with the SPINCOVSE package.

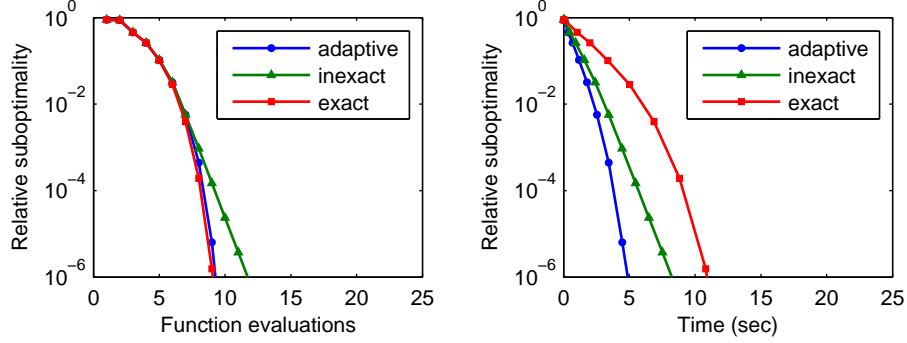


Figure 1: Inverse covariance estimation problem (Estrogen dataset). Convergence behavior of proximal BFGS method with three subproblem stopping conditions.

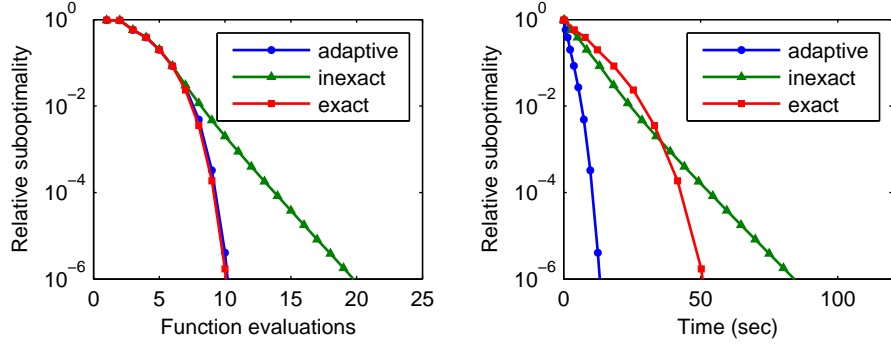


Figure 2: Inverse covariance estimation problem (Leukemia dataset). Convergence behavior of proximal BFGS method with three subproblem stopping conditions.

We solve the inverse covariance estimation problem (4.1) using a proximal BFGS method, *i.e.* H_k is updated according to the BFGS updating formula. To explore how inexact search directions affect the convergence behavior, we use three rules to decide how accurately to solve subproblem (2.5):

1. adaptive: stop when the adaptive stopping condition (2.21) is satisfied;
2. exact: solve subproblem exactly;
3. inexact: stop after 10 iterations.

We plot relative suboptimality versus function evaluations and time on the Estrogen dataset in Figure 1 and the Leukemia dataset in Figure 2.

On both datasets, the exact stopping condition yields the fastest convergence (ignoring computational expense per step), followed closely by the adaptive stopping condition (see Figure 1 and 2). If we account for time per step,

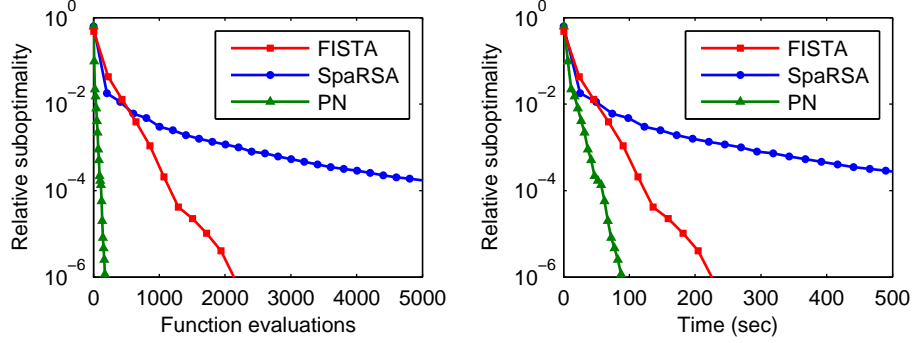


Figure 3: Logistic regression problem (**gisette** dataset). Proximal L-BFGS method ($L = 50$) versus FISTA and SpaRSA.

then the adaptive stopping condition yields the fastest convergence. Note that the adaptive stopping condition yields superlinear convergence (like the exact proximal BFGS method). The third (inexact) stopping condition yields only linear convergence (like a first-order method), and its convergence rate is affected by the condition number of $\hat{\Theta}$. On the Leukemia dataset, the condition number is worse and the convergence is slower.

4.2 Logistic regression

Suppose we are given samples $x^{(1)}, \dots, x^{(n)}$ with labels $y^{(1)}, \dots, y^{(n)} \in \{0, 1\}$. We fit a logit model to our data:

$$\underset{w \in \mathbf{R}^n}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(-y_i w^T x_i)) + \lambda \|w\|_1. \quad (4.2)$$

Again, the regularization term $\|w\|_1$ promotes sparse solutions and λ balances goodness-of-fit and sparsity.

We use two datasets: (i) **gisette**, a handwritten digits dataset from the NIPS 2003 feature selection challenge ($n = 5000$), and (ii) **rcv1**, an archive of categorized news stories from Reuters ($n = 47,000$).² The features of **gisette** have been scaled to be within the interval $[-1, 1]$, and those of **rcv1** have been scaled to be unit vectors. λ was chosen to match the value reported in [28], where it was chosen by five-fold cross validation on the training set.

We compare a proximal L-BFGS method with SpaRSA and the TFOCS implementation of FISTA (also Nesterov’s 1983 method) on problem (4.2). We plot relative suboptimality versus function evaluations and time on the **gisette** dataset in Figure 3 and on the **rcv1** dataset in Figure 4.

The smooth part requires many expensive exp/log operations to evaluate. On the dense **gisette** dataset (30 million nonzero entries in a 6000 by 5000

²These datasets are available from <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets>.

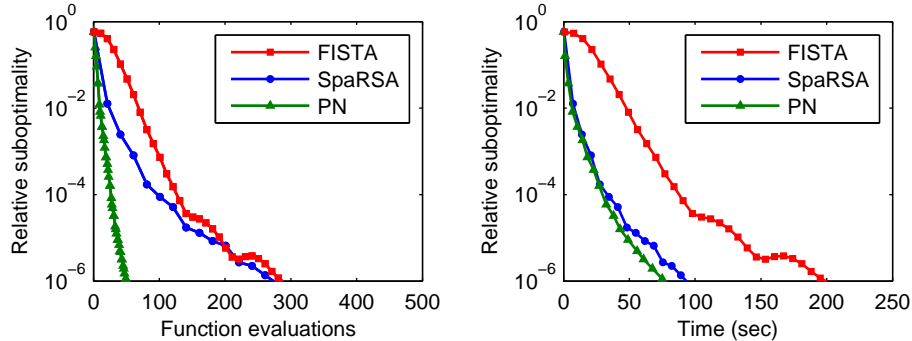


Figure 4: Logistic regression problem (`rcv1` dataset). Proximal L-BFGS method ($L = 50$) versus FISTA and SpaRSA.

design matrix), evaluating g dominates the computational cost. The proximal L-BFGS method clearly outperforms the other methods because the computational expense is shifted to solving the subproblems, whose objective functions are cheap to evaluate (see Figure 3). On the sparse `rcv1` dataset (40 million nonzero entries in a 542,000 by 47,000 design matrix), the cost of evaluating g makes up a smaller portion of the total cost, and the proximal L-BFGS method barely outperforms SpaRSA (see Figure 4).

5 Conclusion

Given the popularity of first-order methods for minimizing composite functions, there has been a flurry of activity around the development of Newton-type methods for minimizing composite functions [12, 2, 16]. We analyze proximal Newton-type methods for such functions and show that they have several strong advantages over first-order methods:

1. They converge rapidly near the optimal solution, and can produce a solution of high accuracy.
2. They are insensitive to the choice of coordinate system and to the condition number of the level sets of the objective.
3. They scale well with problem size.

The main disadvantage is the cost of solving the subproblem. We have shown that it is possible to reduce the cost and retain the fast convergence rate by solving the subproblems inexactly. We hope our results kindle further interest in proximal Newton-type methods as an alternative to first-order and interior point methods for minimizing composite functions.

Acknowledgements

We thank Santiago Akle, Trevor Hastie, Nick Henderson, Qiang Liu, Ernest Ryu, Ed Schmerling, Carlos Sing-Long, Walter Murray, and three anonymous referees for their insightful comments. J. Lee was supported by a National Defense Science and Engineering Graduate Fellowship (NDSEG) and an NSF Graduate Fellowship. Y. Sun and M. Saunders were partially supported by the DOE through the Scientific Discovery through Advanced Computing program, grant DE-FG02-09ER25917, and by the NIH, award number 1U01GM102098-01. M. Saunders was also partially supported by the ONR, grant N00014-11-1-0067.

A Proofs

Lemma 3.4. *Suppose g is twice continuously differentiable and $\nabla^2 g$ is locally Lipschitz continuous with constant L_2 . If $\{H_k\}$ satisfy the Dennis-Moré criterion (3.1) and their eigenvalues are bounded, then the unit step length satisfies the sufficient descent condition (2.16) after sufficiently many iterations.*

Proof. Since $\nabla^2 g$ is locally Lipschitz continuous with constant L_2 ,

$$g(x + \Delta x) \leq g(x) + \nabla g(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 g(x) \Delta x + \frac{L_2}{6} \|\Delta x\|^3.$$

We add $h(x + \Delta x)$ to both sides to obtain

$$\begin{aligned} f(x + \Delta x) &\leq g(x) + \nabla g(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 g(x) \Delta x \\ &\quad + \frac{L_2}{6} \|\Delta x\|^3 + h(x + \Delta x). \end{aligned}$$

We then add and subtract $h(x)$ from the right-hand side to obtain

$$\begin{aligned} f(x + \Delta x) &\leq g(x) + h(x) + \nabla g(x)^T \Delta x + h(x + \Delta x) - h(x) \\ &\quad + \frac{1}{2} \Delta x^T \nabla^2 g(x) \Delta x + \frac{L_2}{6} \|\Delta x\|^3 \\ &\leq f(x) + \Delta + \frac{1}{2} \Delta x^T \nabla^2 g(x) \Delta x + \frac{L_2}{6} \|\Delta x\|^3 \\ &\leq f(x) + \Delta + \frac{1}{2} \Delta x^T \nabla^2 g(x) \Delta x + \frac{L_2}{6m} \|\Delta x\| \Delta, \end{aligned}$$

where we use (2.11). We add and subtract $\frac{1}{2}\Delta x^T H \Delta x$ to yield

$$\begin{aligned}
f(x + \Delta x) &\leq f(x) + \Delta + \frac{1}{2}\Delta x^T (\nabla^2 g(x) - H) \Delta x \\
&\quad + \frac{1}{2}\Delta x^T H \Delta x + \frac{L_2}{6m} \|\Delta x\| \Delta \\
&\leq f(x) + \Delta + \frac{1}{2}\Delta x^T (\nabla^2 g(x) - H) \Delta x \\
&\quad - \frac{1}{2}\Delta + \frac{L_2}{6m} \|\Delta x\| \Delta,
\end{aligned} \tag{A.1}$$

where we again use (2.11). $\nabla^2 g$ is locally Lipschitz continuous and Δx satisfies the Dennis-Moré criterion. Thus,

$$\begin{aligned}
&\frac{1}{2}\Delta x^T (\nabla^2 g(x) - H) \Delta x \\
&= \frac{1}{2}\Delta x^T (\nabla^2 g(x) - \nabla^2 g(x^*)) \Delta x + \frac{1}{2}\Delta x^T (\nabla^2 g(x^*) - H) \Delta x \\
&\leq \frac{1}{2} \|\nabla^2 g(x) - \nabla^2 g(x^*)\| \|\Delta x\|^2 + \frac{1}{2} \|(\nabla^2 g(x^*) - H) \Delta x\| \|\Delta x\| \\
&\leq \frac{L_2}{2} \|x - x^*\| \|\Delta x\|^2 + o(\|\Delta x\|^2).
\end{aligned}$$

We substitute this expression into (A.1) and rearrange to obtain

$$f(x + \Delta x) \leq f(x) + \frac{1}{2}\Delta + o(\|\Delta x\|^2) + \frac{L_2}{6m} \|\Delta x\| \Delta.$$

We can show Δx_k converges to zero via the argument used in the proof of Theorem 3.1. Hence, for k sufficiently large, $f(x_k + \Delta x_k) - f(x_k) \leq \frac{1}{2}\Delta_k$. \square

References

- [1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.
- [2] S. Becker and M.J. Fadili. A quasi-Newton proximal splitting method. In *Adv. Neural Inf. Process. Syst. (NIPS)*, 2012.
- [3] S.R. Becker, E.J. Candès, and M.C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Math. Prog. Comp.*, 3(3):165–218, 2011.
- [4] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.
- [5] E.J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

- [6] R.S. Dembo, S.C. Eisenstat, and T. Steihaug. Inexact Newton methods. *SIAM J. Num. Anal.*, 19(2):400–408, 1982.
- [7] J.E. Dennis and J.J. Moré. A characterization of superlinear convergence and its application to quasi-Newton methods. *Math. Comp.*, 28(126):549–560, 1974.
- [8] S.C. Eisenstat and H.F. Walker. Choosing the forcing terms in an inexact Newton method. *SIAM J. Sci. Comput.*, 17(1):16–32, 1996.
- [9] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007.
- [10] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostat.*, 9(3):432–441, 2008.
- [11] M. Fukushima and H. Mine. A generalized proximal point algorithm for certain non-convex minimization problems. *Internat. J. Systems Sci.*, 12(8):989–1000, 1981.
- [12] C.J. Hsieh, M.A. Sustik, I.S. Dhillon, and P. Ravikumar. Sparse inverse covariance matrix estimation using quadratic approximation. In *Adv. Neural Inf. Process. Syst. (NIPS)*, 2011.
- [13] D. Kim, S. Sra, and I.S. Dhillon. A scalable trust-region algorithm with application to mixed-norm regression. In *Int. Conf. Mach. Learn. (ICML)*. Citeseer, 2010.
- [14] Z. Lu and Y. Zhang. An augmented Lagrangian approach for sparse principal component analysis. *Math. Prog. Ser. A*, pages 1–45, 2011.
- [15] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer, 2003.
- [16] P.A. Olsen, F. Oztoprak, J. Nocedal, and S.J. Rennie. Newton-like methods for sparse inverse covariance estimation. In *Adv. Neural Inf. Process. Syst. (NIPS)*, 2012.
- [17] M. Patriksson. Cost approximation: a unified framework of descent algorithms for nonlinear programs. *SIAM J. Optim.*, 8(2):561–582, 1998.
- [18] M. Patriksson. *Nonlinear Programming and Variational Inequality Problems: A Unified Approach*. Kluwer Academic Publishers, 1999.
- [19] B. Rolfs, B. Rajaratnam, D. Guillot, I. Wong, and A. Maleki. Iterative thresholding algorithm for sparse inverse covariance estimation. In *Adv. Neural Inf. Process. Syst. (NIPS)*, pages 1583–1591, 2012.
- [20] M. Schmidt. *Graphical model structure learning with ℓ_1 -regularization*. PhD thesis, University of British Columbia, 2010.

- [21] M. Schmidt, D. Kim, and S. Sra. Projected Newton-type methods in machine learning. In S. Sra, S. Nowozin, and S.J. Wright, editors, *Optimization for Machine Learning*. MIT Press, 2011.
- [22] M. Schmidt, E. Van Den Berg, M.P. Friedlander, and K. Murphy. Optimizing costly functions with simple constraints: A limited-memory projected quasi-Newton algorithm. In *Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2009.
- [23] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, pages 267–288, 1996.
- [24] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *unpublished*, 2009.
- [25] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Math. Prog. Ser. B*, 117(1):387–423, 2009.
- [26] S.J. Wright, R.D. Nowak, and M.A.T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Trans. Signal Process.*, 57(7):2479–2493, 2009.
- [27] J. Yu, S.V.N. Vishwanathan, S. Günter, and N.N. Schraudolph. A quasi-Newton approach to nonsmooth convex optimization problems in machine learning. *J. Mach. Learn. Res.*, 11:1145–1200, 2010.
- [28] G.X. Yuan, C.H. Ho, and C.J. Lin. An improved glmnet for ℓ_1 -regularized logistic regression. *J. Mach. Learn. Res.*, 13:1999–2030, 2012.